STAT 5443, LITERATURE REVIEW

Ariel Mundo

2020-12-10

PART 2

LITERATURE REVIEW

Finding Online Extremists in Social Networks

Klausen et. al.

https://arxiv.org/pdf/1610.06242.pdf

The paper by Klausen et al covers the use of machine learning algorithms to detect extremists on social media using behavioral and network features.The study collects large amounts of data from users from Twitter that have been identified that have some interaction with the extremist group ISIS (The Islamic State of Iraq and the Levant). The authors collected 1.3 million user profiles related to ISIS and their friends and followers, and 4.8 million tweets. Due to the large amount of collected data, it is clear that a machine learning approach is appropriate to analyze the data.

The first model the authors build is to use logistic regression to predict account suspensions using a dataset of 5,000 accounts. The predictors in this case were network and account features including followed accounts known to belong to ISIS, date and time, geo-location and others. It seems interesting to me the use of mixed variables (binary and numeric) to fit a logistic model.The amount of predictors was large, with 2,367 binary variables assigned to the status of the account following another ISIS account and 7 other predictors related to the account information. Their results suggest that using a ℓ_1 regularization the model can detect about 60% of the suspended users with a 10% false positivity rate. Misclassified accounts were identified as having few tweets or no posting content at all. The regularization constant was identified as optimal at 10, and the model assigned non-zero coefficients to 89 of the predictors. From the original 2000+ predictors, this indicates a significant reduction, and since 81 out of the 89 predictors are related to account following, this means that predictors associated with user-to-user interactions are the strongest to predict suspension. This makes sense, as the behavior of the user and his interactions are likely to indicate if he is going to be suspended by engaging with content associated to ISISI accounts. However, the percentage of detection seems to not be too high in this case (60%), and the authors provide no direct explanation of the implications in a real-life scenario where only such percentage is detected.

The next model of supervised machine learning approach was developed to detect if two user profiles belong to the same user (multiple accounts).The similarity metrics used were:

-Screen name: Distance between two strings to identify similarly named accounts was used (The Levenshtein ratio). -Similarity between profile pictures using matching shade patterns. The model used 3,944 profiles and the classification was specified as 1 if profiles belonged to the same user and 0 if the they belonged to different users. A logistic regression with a ℓ_1 regularization with $\lambda = 10$ identified as the best regularization parameter using cross validation. Interestingly, feature with the highest regression coefficient was the Levenshtein ratio (7.05). In other words, the predictor with the highest regression coefficient is the one that relates to similarities in user name. It was expected that the classifier correctly identifies 80% of accounts pair belonging to the same user and that it would misclassify % of the different user pairs.

A refollowing model (to determine how users reconnect) was also developed. The model would predict the probability that a user will re-follow the same “friend” upon this person opens a new Twitter account. Here the data was clustered and the features were set as User0 (from a suspended user account) and Friend, the account that was to be followed.

The model used in this case was a quadratic kernel logistic regression, to accomodate interactions between the terms, allowing to linear fit in all terms. The data was divided between training, validation and test sets (50%, 25%, 25%). The regularization term λ was selected as 10^{-5} using the AUC criterion instead of cross validation. One major limitation of the quadratic kernel approach is that the expressions on the fit models are not easy to interpret, and the AUC approach for the model resulted in 0.66. The authors suggest that this indicates that there is a refollowing behavior that is beyond the users in the dataset.

The final model was developed to find the new account of a previously suspended user.This the most extensive section of the paper and probably the most important contribution of the authors to the field. They develop theorems to perform incorporate a stochastic search as the core of the model. This included derivations for conditional reconnection and a *policy cost*, or the cost of performing the search for the new account of a previously suspended user. This model was developed and tested on a much smaller number of accounts (169 pairs), where 15 were used for testing. Of key importance for this model was the probability for conditional existence $\rho(t)$, which uses a Bayesian approach. It was used with different policies as criteria on when the analysis would terminate.

In this case, the policies were: Optimal, Greedy, Min-N, Max-P. The most relevant policies where the Optimal and Greedy approaches. In the first case, the probability of finding a new account at each stage is maximized, and in the second the expected cost is minimized. These were the more relevant because the analysis showed that the costs for both approaches were very similar in most cases, which would indicate a minimization of cost and a maximizing the probability at each stage have the same effect.

One thing that stands out is that in all cases the regularization parameter λ was found to be 10 for the best performance in cross-validation. While the authors do not indicated any specifics on the reasons behind this, it would be interesting to see if other papers published in the same research topic have regularization values that are close to this value. If that were the case, it would indicate that a smaller set of predictors can be used in the classification models, or that there is an interesting behavior on the behavior of social networks that makes this regularization parameter to have a constant value.

Finally, the authors indicated that their models can be used beyond Twitter as the algorithms used behavioral and statistical approaches but are were not limited by a specific type of social network. It would be interesting to see how these models perform with models that have been developed specifically for each one of the major social networks (Instagram, Facebook and Twitter) and what predictors these specific models take into account to make the predictions. Overall I think the topic is quite interesting, as an user-based policing in social networks for extremists is certainly impossible in a world where billions of persons have accounts and each social networks has millions of features that can be exploited. This paper made me realize uses of statistical methods of machine learning to address this technological need.